Effects of Missing Data on the Parameters of Multiple Regression Model (MRM)

Etaga Harrison. O*, Ngonadi Lilian, Aforka Kenechukwu F and Etaga Njideka C

Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

*Corresponding author

Received: 27 January 2023; Revised: 15 February 2023; Accepted: 21 February 2023; Published: 21 March 2023

Abstract: Multiple Regression Models are used in prediction the nature of relationship between one dependent variable and more than more independent variables. There are so many assumptions the guide the estimation of the parameters of the model. The interpretations of parameters are always subjected to the nature of data involved. Missing values tends to limit the fullness of information in analysis. It is therefore necessary to check for the effect of missing data on the parameters of the Multiple Regression Model. Data were simulated using Binomial, Geometric, Normal and Exponential Distribution. The simulation was done at different sample sizes of 15, 25, 50 and 100. The level of missingness was moderated at 5%, 10%, 25% and 35%. Two methods of handling missing data were employed, listwise deletion and Mean imputation. Data were analysis using multiple regression and Analysis of Variance. The results shows that the least Mean Square Error (MSE) were obtained at different level of missingness depending of the distribution. There was a significant effect on the parameters of the multiple Regression base on sample sizes.

Keywords: Multiple, Regression, Model, Missing data, MSE

I. Introduction

Researchers have faced the problem of missing quantitative data at some point in the work. Missing data can occur if research informants refuse or forgot to answer a survey question or there might be lost of files as well as data might not be recorded properly. Given the high cost of collecting data, there cannot be wastage of effort of starting all over or to wait until soundproof methods of collecting information are developed. In statistics, Missing data/value is an occurrence when there is no data value stored for the variable in an observation.

Regression analysis is the study of the nature of relationship between dependent variable(s) and independent variable(s). The Simple Linear regression involves just one dependent and one independent variable. The situation where there exist one dependent variable and more than one independent variable is referred to as Multiple Regression (MR). When estimating the parameters of the Multiple Regression, Least Squares Method (LSM) is used most often. There are various factors that can affect the signs or magnitudes of the parameter(s), one of such is that of missing data. There is need to adequately address the problem of missing data before analysis the data to avoid reaching wrong conclusions.

When treating missing data, the most common method and the easiest to apply is the use of only those cases with complete information. An alternative to complete case analysis, there is the use of the mean as a replacement of the missing value. More recently, there are methods that are based on distributional models for the data (such as maximum likelihood and multiple imputation).

Methods of analyzing missing data require assumptions about the nature of the data and about the reasons for the missing observations that are not often acknowledged. There is need to carefully considered the required assumptions before treating missing data. Missing data can lead to problems that affect the interpretation and inference of research results, the understanding and explanation of conclusions made, the strength of the study design, the validity of conclusions about the relationship between variables and may limit the representativeness of the sample.

Avoiding missing data is the optimal means of handling incomplete observations. During data collection phase, the researcher has the opportunity to make decisions about what data to collect, and how to monitor data collection. The scale and distribution of the variables in the data and the reasons for missing data are two critical issues for applying the appropriate missing data techniques. This paper there seek to evaluate the effects of missing data on the parameter estimates of the Multiple Regression.

II. Literature

There are various forms of missingness. Rubin (1976) introduced the term “Missing Completely At Random” (MCAR). MCAR is a situation where the events that led to the particular data item being missing are independent both of observable variables and of unobservable parameters of interest and occur entirely at random. There is another missingness referred to as Missing AT Random (MAR), which occurs when the missingness is not random, but where missingness can be fully accounted for by variables where there exists complete information. This is when the probability that the responses that are missing depends on the set of observed responses, but not related to the specific missing values.

Sunbul (2018) opined in his study aimed at investigating the impact of different missing data handling methods on model parameter estimation and classification accuracy. Simulated data were used and the data were generated by manipulating the number of items and sample size. In the generated data, two different missing data mechanisms (Missing Completely At Random and Missing At Random) were created according to three different amounts of missing data. The generated missing data was completed by using methods of treating missing data as incorrect, person mean imputation, two-way imputation, and expectation-maximization algorithm imputation.

Graham (2009) noted that most missing data are due to survey non-response, which can vary from an intentional decision (discarding a survey or skipping sensitive items) to a rather more complex reason(s). The problem of missing data is relatively common in almost all researches and can have significant effects on the conclusions that can be drawn from the data.

Pigott (2001) reviewed methods of handling missing data in a research study. He observed that many researchers use ad hoc methods such as complete case analysis, available case analysis (pairwise deletion), or single-value imputation. Though these methods are easily implemented, they require assumptions about the data that rarely hold in practice. Model-based methods such as maximum likelihood using the EM algorithm and multiple imputation hold more promise for dealing with difficulties caused by missing data.

Rubin (1976) introduced the term “Missing Completely At Random “(MCAR) to describe data where the complete cases are a random sample of the originally identified set of cases.

Little and Rubin, Schafer (1997) discuss methods that can be used for non-ignorable missing data. They observed that ruling out a non-ignorable response mechanism can simplify survey items.

Schafer (1997) reports on simulation studies that provide evidence of the robustness with missing data and using missing variable code as a predictor in a regression model. Penny et al (2012 stated in their book that the simplest approach to missing data, and the one that is the default in virtually all statistical packages is the method known to statistician as complete case analysis but more commonly known among social scientist as listwise deletion, in this method, cases are deleted from the sample if they have missing data on any of the variables in the analysis to be conducted.

Peter et al (2015) observed that Missing data are part of almost all researches and introduce an element of ambiguity into data analysis. It follows that there is need to consider them appropriately in order to provide an efficient and valid analysis. they compared 6 different imputation methods: Mean, K-nearest neighbors (KNN), fuzzy K-means (FKM), singular value decomposition (SVD), Bayesian principal component analysis (bPCA) and multiple imputations by chained equations (MICE). Comparison was performed on four real datasets of various sizes (from 4 to 65 variables), under a missing completely at random (MCAR) assumption, and based on four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time. Our results suggest that bPCA and FKM are two imputation methods of interest which deserve further consideration in practice.

Nakai and Weiming (2011), observed that in well-controlled situations, missing data always occur in longitudinal data analysis. Missing data may degrade the performance of confidence intervals, reduce statistical power and bias parameter estimate. They review and discuss general approaches for handling miss data in longitudinal studies. They started by first illustrating the mechanism of missing data. Then focused on the methods for handling missing values in longitudinal data analysis. In the end, they summarized and discussed the characteristics of each method

Howell (2000) observed that the treatment of missing data has been an issue in statistics for some time, but it has come to the fore in recent years. He noted that the current interest in missing data stems mostly from the problems caused in surveys and census data. He suggested several methods of treating missing values and elaborately gave some examples of how to deal with missingness.

Allison (2001) uses the example of ‘missingness’ for data on income being dependent on marital status. Perhaps unmarried couples are less likely to report their income than married ones. Unmarried couples probably have lower incomes than married ones, and it would at first appear that missingness on income is related to the value of income itself. But the data would still be MAR if the conditional probability of missingness were unrelated to the value of income within each marital category. Here the real question is whether the value of the dependent variable determines the probability that it will be reported, or whether there is some other variable (X) where the probability of missingness on Y is conditional on the levels of X. To put it more formally, data are MAR if $p(Y \text{ missing } |Y,X) = p(Y \text{ missing } | X)$.

Dempster, Laird and Rubin (1977). Although, this method was considered to be more superior to the previous adhoc methods in that they produced better estimates with smaller and acceptable standard errors. Finally, in the late 80’s, more superior methods such as Multiple Imputation, were developed. These methods were proved to be flexible and produced smaller standard errors as compared to earlier methods.

Cohen and Cohen (1983), have suggested that when the missing data is on the dependent variable, the subject may be dropped from the analysis. However, if the missing data is among the independent variables, it might be intrusive to determine what proportion of the data is missing.

Truxillo (2005) has suggested that the EM covariance matrix and vector of means can also be used as input for procedures that entail inference, but that one must then use a nominal sample size that properly accounts for the fact that some data are missing.

In Little's test of MCAR (Little 1988), the data y_i, (i = 1, 2,...,n) are modeled as p-dimensional multivariate normal with mean vector μ and covariance matrix Σ, with part of the components in y_i's missing. When the normality is not satisfied, Little's test still works in the asymptotic sense for quantitative random vectors y_i's but is not suitable for categorical variables

De Silva et al (2007) observed that the presence of missing values in rainfall data is a common problem in the process of data analysis. They suggested ways of treating missing data in temperature and rainfall data so as to obtain the best method.

III. Method

Data for this study were simulated from four different distributions. Two discrete and two continuous distributions were considered. The distributions considered are Binomial, Geometric, Normal and Exponential distributions. The following specifications were used

i. Binomial (p= 0.25, n=15, 25, 50, 100)
ii. Geometric (p = 0.5, n=15, 25, 50, 100)
iii. Normal (μ = 9.7, σ = 1.3, n=15, 25, 50, 100)
iv. Exponential (μ = 6.9, n=15, 25, 50, 100)

Level of missingness considered

The computations and analysis were done for distributions at various level of missingness. The level of missingness considered varied from 5% to 35%. To be specific, the different level of missingness are; 5%, 10%, 25% and 35%.

Missing Data Method considered.

Two methods of handling missing data were considered in other to examine their effect on the parameters of a Multiple Regression Model (MRM). The two methods considered are Mean Imputation and listwise deletion methods

Multiple Regression Model.

Regression Models are statistical procedures that deals with the nature of relationship between dependent variable(s) and Independent Variable(s). Multiple regression model that deals with one dependent and more than one independent variables.

Model specification

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_n x_n + e_i \quad . \quad . \quad . \quad . \quad . \qquad (1)$$

Where

Y_1 is the dependent variable

B_0 is the intercept

β_i are the regression coefficients that represent the changes in y relative to a ne – unit change in x_i's

e_i is the residual term.

The parameters of the model β_i ae estimated using

$$\beta = (X'X)^{-1}(X'Y) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (2)$$

Where

$$(X'X)^{-1} = \begin{bmatrix} n & \sum_{i=1}^{n} X_1 & \sum_{i=1}^{n} X_2 & \cdots & \sum_{i=1}^{n} X_n \\ \sum_{i=1}^{n} X_1 & \sum_{i=1}^{n} X^2{}_1 & \sum_{i=1}^{n} X_i X_2 & \cdots & \sum_{i=1}^{n} X_1 X_n \\ . & . & . & & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ \sum_{i=1}^{n} X_n & \sum_{i=1}^{n} X_n X_1 & \sum_{i=1}^{n} X_n X_2 & ::: & \sum_{i=1}^{n} X^2{}_n \end{bmatrix}$$

$$(X'Y) = \begin{bmatrix} \sum_{i=1}^{n} Y \\ \sum_{i=1}^{n} X_1 Y \\ \vdots \\ . \\ \sum_{i=1}^{n} X_n Y \end{bmatrix}$$

Model Performance

The performances of the MR model were determined using Regression Analysis of Variance ANOVA given in Table 1.

Table 1: Regression ANOVA for model adequacy

Source of Variation	Dere of Freedom	Sum of Squares	Mean Squares (MS)	F-Ratio
Regression	k-1	SSR	$MSR = \frac{SSR}{K-1}$	$F = \frac{MSR}{MSE}$
Error	N-K	SSE	$MSE = \frac{SSE}{N-K}$	
Total	N-1	SST		

$$Sum\ of\ Squares\ of\ Regression\ (SSR) = \beta' X' Y - n\bar{Y}^2 \quad . \quad . \quad . \qquad 3$$

$$Sum\ of\ Squares\ of\ Error\ (SSE) = Y'Y - \beta' X' Y \ . \quad . \quad . \quad . \qquad 4$$

$$Sum\ of\ Squares\ of\ Total\ (SST) = Y'Y - n\bar{Y}^2 \quad . \quad . \quad . \quad . \quad . \qquad 5$$

Mean Sum of Squares Regression (MSR)

$$MSR = \frac{SSR}{K-1} . \quad . \quad . \quad . \quad . \quad . \quad .6$$

Mean Sum of Squares Error (MSE)

$$MSE = \frac{SSE}{N-K} \quad . \quad . \quad . \quad . \quad . \quad . \quad 7$$

IV. Results

Multiple regression models were fitted for the data simulated at different sample sizes. The F-Values, MSR, MSE and P-Values where obtained. The model with the lowest MSE is considered as the best.

Table 2: Estimation Results for Complete Data Sets by Distribution

Sample size	Distribution	B_0	B_1	B_2	F-Values	MSR	MSE	P-Value
15	Binomial	7.69	-0.435	-0.073	0.62	2.9360	4.7107	0.553
25		4.71	-0.056	0.131	0.18	0.7702	4.2863	0.837
50		5.168	-0.067	-0.029	1.14	0.3655	2.5908	0.869
100		5.224	0.0038	-0.017	0.02	0.0643	3.6630	0.983
15	Geometric	1.59	0.741	-0.362	3.28	15.010	4.576	0.075
25		2.57	-0.133	0.092	0.37	1.7995	4.8437	0.694
50		2.887	0.016	-0.194	1.21	4.11227	3.40629	0.308
100		2.618	-0.014	-0.070	0.35	0.91717	2.60336	0.704
15	Normal	4.96	0.358	0.182	1.03	1.7354	1.6896	0.387
25		9.08	0.121	-0.042	0.17	0.21985	1.26686	0.842
50		9.12	0.101	-0.023	0.28	0.44113	1.56528	0.756
100		8.58	0.0238	0.083	0.38	0.56528	1.49900	0.687
15	Exponential	6.29	0.034	-0.105	0.10	59.08	58.573	0.905
25		8.05	-0.281	0.106	0.97	46.38	47.83	0.395
50		7.74	-0.184	-0.034	1.23	62.820	51.223	0.303
100		6.22	-0.128	0.0108	0.83	31.937	38.5476	0.440

For the Complete data set, the Parameters of the Multiple Regression were estimated for all the distributions and their MSE obtained. The results showed that the Normal distribution at the sample size of 25 had the least MSE of 1.26686. It was also noticed that at all sample sizes, the Normal distribution had the least MSE. This buttressed the fact that some assumptions guiding Multiple regression has some link to Normal distribution.

Table 3: Estimation Results for Binomial Data Sets by degree of missingness and method of handling

Level	Method	Sample Size	β_0	β_1	β_2	F-Values	MSR	MSE	P-Value
5%	Listwise	12	9.89	-0.775	-0.246	1.94	7.832	4.0373	0.199
		22	3.99	-0.063	0.243	0.5	2.2406	4.4914	0.615
		41	4.25	0.104	-0.029	0.23	0.58254	2.4892	0.792
		85	4.995	0.0187	0.035	0.06	0.2167	3.8277	0.945
	Mean Imputation	15	8.91	-0.617	-0.19	2.3	6.8138	3.3644	0.175
		25	4.47	-0.064	0.178	0.32	1.322	4.1916	0.733
		50	4.518	0.068	-0.021	0.12	0.27552	2.26351	0.886
		100	5.199	0.0052	-0.016	0.02	0.06112	3.60634	0.983
10%	Listwise	10	7.01	-0.387	-0.0044	0.32	1.93746	6.01787	0.735
		18	4.83	-0.095	0.167	0.25	0.913	3.6449	0.782
		37	5.11	-0.079	-0.025	0.11	0.35105	3.0652	0.892
		71	5.023	-0.018	0.056	0.14	0.58401	4.05987	0.866

	Mean Imputation	15	7.12	-0.378	0.064	0.53	2.1816	4.1364	0.603
		25	5.67	-0.154	0.054	0.29	1.003	3.4481	0.75
		50	5.466	-0.082	-0.07	0.31	0.7397	2.378	0.734
		100	5.152	-0.022	0.026	0.08	0.2577	3.2576	0.924
25%	Listwise	3	7.8	0	-0.4	0	1.333	0	
		9	2.55	-0.273	1.182	1.23	7.2727	5.9091	0.357
		17	7.65	-0.149	-0.389	0.8	2.7014	3.3956	0.417
		34	4.75	0.128	0.034	0.38	1.4899	3.9162	0.687
	Mean Imputation	15	7.59	-0.169	-0.241	0.62	1.7215	0.27737	0.554
		25	5.58	-0.26	0.168	0.81	2.865	3.519	0.456
		50	4.94	0.006	-0.004	0	0.00205	1.74229	0.999
		100	5.332	0.0039	-0.0011	0	0.00295	2.52558	0.999
35%	Listwise	3	7.7	0.189	0.62	0.07	1.451	19.841	0.934
		4	4.1	-0.15	0.4		1		
		13	7.61	-0.3	-0.143	0.92	1.5774	1.7153	0.43
		22	5.4	-0.18	0.109	0.52	2.7248	5.2371	0.603
	Mean Imputation	15	7.66	-0.432	0.101	0.41	1.0207	2.4632	0.67
		25	3.94	0.178	-0.026	0.21	0.51317	2.48744	0.815
		50	5.12	-0.018	-0.037	0.05	0.07132	1.53353	0.955
		100	5.233	-0.0583	0.059	0.44	1.173	2.6775	0.647

For the Binomial data set, the Parameters of the Multiple Regression were also estimated for all the distributions and their MSE also obtained. It was observed that when listwise deletion was used, the least MSE of 1.7153 was obtained at 35% level of Missingness. When mean imputation was used, the least MSE of 1.53353 was obtained at 35% level of Missingness. Overall, The Mean Imputation method gave the least MSE of 1.53353 for the Binomial Distribution. This shows that missingness of data can affect the significant of the model.

Table 4: Estimation Results for Geometric Data Sets by degree of missingness and method of handling

Level	Method	Sample Size	β_0	β_1	β_2	F-Values	MSR	MSE	P-Value
5%	Listwise	12	1.25	0.845	-0.452	0.2	1.0964	5.4277	0.819
		22	2.31	-0.071	0.1	0.2	1.0964	5.4277	0.819
		41	2.269	-0.03	-0.089	0.76	0.9951	1.3135	0.476
		86	2.669	-0.075	-0.053	0.35	0.7827	2.2473	0.707
	Mean Imputation	15	1.75	0.77	-0.447	4.24	16.638	3.923	0.04
		25	2.31	-0.77	0.116	0.31	1.5317	4.8541	0.734
		50	2.759	-0.084	-0.104	0.7	2.015	2.889	0.503
		100	2.574	-0.078	-0.0687	0.41	0.86718	2.09346	0.662
10%	Listwise	9	0.8	-0.0048	-1.082	6.35	24.971	3.935	0.033
		16	3.6	1.488	0.142	1.33	7.289	5.494	0.299
		36	2.871	0.025	-0.177	0.62	2.5702	4.117	0.542
		73	2.603	-0.024	-0.062	0.2	0.58274	2.91877	0.819

	Mean Imputation	15	1.67	0.895	-0.0837	4.82	18.473	3.831	0.029
		25	2.8	-0.273	0.101	0.7	3.215	4.575	0.506
		50	2.783	0.024	-0.145	0.65	2.1005	3.2137	0.525
		100	2.446	0.036	-0.0562	0.31	0.7529	2.4133	0.733
25%	Listwise	6	-3.18	0.529	4.31	24.66	29.615	1.2	0.014
		8	6.43	-0.669	-0.038	0.78	9.2614	11.8704	0.507
		20	2.904	-0.229	-0.167	0.84	1.74	2.034	0.448
		39	2.205	0.06	-0.018	0.1	0.27252	2.66007	0.903
	Mean Imputation	15	0.88	0.783	0.061	5.09	16.5673	3.2539	0.025
		25	2.44	-0.174	0.163	0.89	0.3927	4.399	0.424
		50	3.126	-0.107	-0.152	0.85	2.599	3.065	0.435
		100	2.177	0.105	-0.0081	0.51	1.1201	2.18309	0.6
35%	Listwise	3	0	2	1		2.33		
		4	2.32	0.113	-0.03318	0.18	0.8784	4.9932	0.86
		10	1.61	0.325	0.174	1.23	3.8943	3.1931	0.349
		29	2.504	-0.127	-0.005	0.39	0.67428	1.72267	0.68
	Mean Imputation	15	3.51	-0.153	0.484	0.24	1.183	4.8778	0.788
		25	1.79	-0.02	0.165	0.21	0.79883	3.85636	0.814
		50	2.938	0.073	-0.214	1.67	4.365	2.6201	0.2
		100	2.496	0.002	-0.06	0.18	0.32027	1.77833	0.835

For the Geometric data set, the Parameters of the Multiple Regression were also estimated for all the distributions and their MSE also obtained. It was observed that when listwise deletion was used, the least MSE of 1.200 was obtained at 25% level of Missingness. When mean imputation was used, the least MSE of 1.77833 was obtained at 35% level of Missingness. Overall, the listwise gave the lowest MSE of 1.200 at the 25% level of missingness. This shows that missingness of data can affect the significant of the model.

Table 5: Estimation Results for Normal Data Sets by degree of missingness and method of handling

Level	Method	Sample Size	β_0	β_1	β_2	F-Values	MSR	MSE	P-Value
5%	Listwise	12	4.44	0.285	0.308	0.61	1.19	1.16	0.563
		21	9.16	0.231	-0.154	0.42	0.5923	1.4129	0.663
		41	9.85	0.051	-0.049	0.12	0.1758	1.4358	0.885
		87	8.75	-0.0034	0.095	0.41	0.61767	1.51295	0.666
	Mean Imputation	15	7.22	0.225	0.089	0.4	0.706	1.764	0.679
		25	9.34	0.175	-0.122	0.31	0.3816	1.2509	0.74
		50	9.73	0.05	-0.037	0.11	0.1675	1.5335	0.897
		100	8.59	0.0137	0.09	0.41	0.60858	1.47203	0.663
10%	Listwise	9	4.69	0.54	0.052	0.75	1.55829	2.08069	0.512
		16	10.28	0.177	-0.223	0.27	0.4011	1.5009	0.77
		35	9.86	0.132	-0.134	0.82	1.3109	1.6003	0.45

		71	9.15	-0.013	0.061	0.14	0.21896	1.54941	0.868
	Mean Imputation	15	5.79	0.475	-0.02	1.73	210286	1.21738	0.0219
		25	9.61	0.167	-0.154	0.38	0.4456	1.1688	0.687
		50	9.4	0.128	-0.086	0.65	0.9669	1.4905	0.527
		100	8.98	0.0079	0.0571	0.17	0.22786	1.30281	0.84
25%	Listwise	3	-13.3	1.949	0.1591		2.82954		
		10	12.2	-0.09	-0.142	0.26	0.18536	0.71218	0.778
		17	12.5	-0.236	-0.009	0.64	0.95935	1.50272	0.543
		45	8	-0.186	0.356	2.67	4.082	1529	0.081
	Mean Imputation	15	0.31	0.36	0.642	1.71	2.448	1.429	0.221
		25	9.1	0.16	-0.089	0.43	0.4143	0.9717	0.658
		50	11.04	-0.158	0.043	0.66	0.7706	1.1663	0.521
		100	9.23	-0.063	0.091	0.48	0.551	1.1649	0.622
35%	Listwise	3	0.7156	0.9261	0.1097		1.20688		
		4	2.7	0.2066	0.603	14.37	1.9664	0.1369	0.183
		11	13.79	-0.097	-0.383	2.53	1.0266	0.4375	0.158
		22	9.87	0.077	-0.098	0.23	0.2625	1.1521	0.798
	Mean Imputation	15	9.64	0.088	-0.05	0.11	0.17326	1.63866	0.9
		25	8.81	-0.052	0.181	0.94	0.47171	0.5041	0.407
		50	7.59	0.228	-0.019	1.33	1.25975	0.94411	0.273
		100	9.81	0.0227	-0.037	0.1	0.09002	0.90819	0.906

For the Normal data set, the Parameters of the Multiple Regression were also estimated for all the distributions and their MSE also obtained. It was observed that when listwise deletion was used, the least MSE of 0.1365 was obtained at 35% level of Missingness. When mean imputation was used, the least MSE of 0.5041 was obtained at 35% level of Missingness. Overall, the listwise gave the lowest MSE of 0.1365 at the 35% level of missingness. This shows that missingness of data can affect the significant of the model.

Table 6: Estimation Results for Exponential Data Sets by degree of missingness and method of handling.

Level	Method	Sample Size	β_0	β_1	β_2	F-Values	MSR	MSE	P-Value
5%	Listwise	12	6.38	0.021	-0.079	0.04	3.791	76.8096	0.961
		22	8.7	-0.298	0.128	1.04	51.14	50.09	0.372
		41	8.32	-0.201	-0.107	1.17	67.21	57.56	0.322
		85	6.22	-0.124	0.0131	0.63	26.1984	41.9085	0.538
	Mean Imputation	15	6.26	0.019	-0.091	0.07	4.375	58.6456	0.929
		25	8.24	-0.298	0.14	1.25	56.3	45.2	0.307
		50	7.69	-0.184	-0.08	1.14	56.39	49.55	0.327
		100	6.47	-0.142	0.0022	0.95	36.0933	37.9739	0.39
10%	Listwise	9	8.41	-0.039	-0.336	0.17	17.183	100.461	0.847
		17	6.78	-0.446	0.456	3.7	161.7	43.7	0.051
		35	7.15	-0.153	0.001	0.78	32.3982	41.3921	0.466
		73	6.51	-0.037	-0.104	0.52	19.1319	36.9509	0.598

	Mean Imputation	15	6.94	-0.085	-0.058	0.12	6.692	55.317	0.887
		25	7.67	-0.336	0.183	2.02	86.09	42.62	0.157
		50	7.74	-0.175	-0.007	1.08	54.549	50.489	0.349
		100	5.72	-0.013	0.0065	0.01	0.4009	37.4961	0.989
25%	Listwise	5	63.1	-3.939	-12.27	15.75	252.56	16.03	0.06
		11	8.37	-0.566	0.501	2.58	153.22	59.4	0.137
		20	8.49	-0.229	0.041	1.4	69.655	49.843	0.274
		10	11.15	-0.43	-0.207	8.34	33.725	4.043	0.014
	Mean Imputation	15	6.04	0.204	-0.173	0.37	19.67	52.89	0.697
		25	6.87	-0.441	0.434	4.31	144.74	33.61	0.026
		50	8.96	-0.195	-0.011	1.35	60.276	44.658	0.269
		100	5.594	-0.1257	0.0686	1.81	28.649	15.794	0.168
35%	Listwise	5	1.52	-0.1363	0.904	9.3	28.838	3.102	0.097
		6	16.68	-1.49	0.211	2.65	65.45	24.66	0.217
		8	19.4	-1.316	-0.22	0.152	160.67	105.68	0.305
		27	9.24	-0.078	-0.292	0.043	36.145	84.795	0.658
	Mean Imputation	15	7.13	-0.163	0.233	0.29	14.51	49.27	0.75
		25	9.38	-0.367	-0.002	0.41	12.9954	31.3236	0.665
		50	8.86	-0.211	-0.065	2.03	69.05	.33.938	0.142
		100	7.38	-0.1544	-0.1379	2.27	64.521	24.442	0.109

For the Exponential data set, the Parameters of the Multiple Regression were also estimated for all the distributions and their MSE also obtained. It was observed that when listwise deletion was used, the least MSE of 3.102 was obtained at 35% level of Missingness. When mean imputation was used, the least MSE of 15.794 was obtained at 25% level of Missingness. Overall, the listwise gave the lowest MSE of 3.102 at the 35% level of missingness. This shows that missingness of data can affect the significant of the model.

The summary of the results can be seen in Table 7.

Table 7: Summary of Results

SN	Distribution	Level	Method	Least MSE	Least-best MSE	Least-Final MSE
1	**Complete**	**-**	**-**	**1.26686**	**1.26686**	**1.26686**
2	Binomial		Listwise	1.7153		
3	Binomial		Mean Imputation	1.53353	1.53353	
4	Geometric		Listwise	1.2000	1.2000	
5	Geometric		Mean Imputation	1.77833		
6	Normal		Listwise	0.1365		
7	Normal		Mean Imputation	0.5041	0.5041	**0.5041**
8	Exponential		Listwise	3.102	3.102	
9	Exponential		Mean Imputation	15.794		

The results in Table 7 shows that when Binomial and Geometric distributions were compared, Binomial had a Least MSE of 1.53353 under mean imputation as against Geometric distribution with MSE of 1.77833. But when Listwise deletion method was used, Geometric distribution had the least MSE of 1.2000 as against Binomial MSE of 1.7153.

Comparing Binomial and Normal distributions, Normal had a Least MSE of 0.5041 under mean imputation as against Binomial distribution with MSE of 1.53353. Also, when Listwise deletion method was used, Normal distribution had the least MSE of 0.1365 as against Binomial MSE of 1.7153.

When Binomial and Exponential distributions were compared, Binomial had a Least MSE of 1.53353 under mean imputation as against Exponential distribution with MSE of 15.794. While under Listwise deletion method, Binomial distribution had the least MSE of 1.7153 as against Exponential distribution with MSE of 3.102.

Comparing Geometric and Normal distributions, Normal had a Least MSE of 0.5041 under mean imputation as against Geometric distribution with MSE of 1.77833. Also when Listwise deletion method was used, Normal distribution had the least MSE of 0.1365 as against Geometric with MSE of 1.2000.

When Geometric and Exponential distributions were compared, Geometric distribution had a Least MSE of 1.77833 under mean imputation as against Exponential distribution with MSE of 15.794. While under Listwise deletion method, Geometric distribution had the least MSE of 1.2000 as against Exponential distribution with MSE of 3.102.

Comparing Exponential and Normal distributions, Normal had a Least MSE of 0.5041 under mean imputation as against Exponential distribution with MSE of 15.794. Also when Listwise deletion method was used, Normal distribution had the least MSE of 0.1365 as against Exponential distribution with MSE of 3.102.

V. Conclusion

The results of the analysis shows that missing data significantly affect the parameters as well as the model significance for various distributions. More specifically, for the complete data set, the results showed that the Normal distribution at the sample size of 25 had the least MSE of 1.26686. It was also noticed that at all sample sizes, the Normal distribution had the least MSE among the four distributions considered. Exploring the results of the Binomial Distribution, it was observed that when listwise deletion was used, the least MSE of 1.7153 was obtained at 35% level of Missingness. When mean imputation was used, the least MSE of 1.53353 was obtained at 35% level of Missingness. Overall, The Mean Imputation method gave the least MSE of 1.53353 for the Binomial Distribution.

For the Normal Distribution, it was observed that when listwise deletion was used, the least MSE of 0.1365 was obtained at 35% level of Missingness. When mean imputation was used, the least MSE of 0.5041 was obtained at 35% level of Missingness. Overall, the listwise gave the lowest MSE of 0.1365 at the 35% level of missingness. For the Exponential Distribution, it was observed that when listwise deletion was used, the least MSE of 3.102 was obtained at 35% level of Missingness. When mean imputation was used, the least MSE of 15.794 was obtained at 25% level of Missingness. Overall, the listwise gave the lowest MSE of 3.102 at the 35% level of missingness.

From the varying values of least MSE's and the sample size effects, it is very important to handle missing values with care and to adhere to the assumptions (if any) on the use of any method of estimations of model parameters so as to obtain consistent and unbiased results.

Reference

1. Allison, P (2001) Missing Data: Quantitative Application in the Social Science. Thousand Oaks, CA: Sage. Vol. 136.
2. Brigga A., Clark T, Wolstenholme, J, Clarke P (2003) Missing Data. Health Economics 12, 377-392
3. Catherine Truxillo (2005) A Comparison of Missing Data Handling Methods. SAS® Institute Inc, Cary, NC.
4. Cohen J and Cohen P (1975): Applied Multiple Regression and Correlation Analysis for the Behavioural Sciences. New York : John Wiley
5. Dempster A. P, Laird, N. M and Rubin D. B. (1977): Maximum Likelihood Estimation from Incomplete data via the EM algorithm. Journal of the Royal Statistical Association B39, 1-38
6. Graham, J. W (2009) Missing data analysis: making it work in the real world. Annu Rev Psychol 60, 549 – 576.
7. Jones, M. P (1996) Indicator and Stratification methods for missing explanatory variables in Multiple Linear Regression. Journal of the American Statistical Association, 91, 222-230
8. Kajornrit, Jesada; Wong, Kok Wai; Fung, Chun Che (2012) A comparative Analysis of soft computing techniques used to estimate missing precipitation records. 9th Biennial Conference of the International Telecommunications Society (ITS): Bangkok, Thailand.

9. Little, R. J. A (1988). A test of missing completely at Random or multivariate data with missing values. Journal of the American Statistical Association 83, 1198 – 1202.
10. Little, R. J. A (1992) Regression with Missing X's: A review. Journal of the American Statistical Association 87, 1227 – 1237.
11. Nakai M and Weiming Ke (2011) Review of methods for handling Missing Data in Longitudinal Data analysis. Int journal of Math. Analysis. Vol 5, no 1. 1-13.
12. De Silva R. P., Dayawansa and Ratnasiri (2007). A comparison of method used n Estimating missing Rainfall data. The Journal of Agricultural Sciences, 2007, Vol 3. No 2.
13. Rubin D. B, (1976) Inference and missing data. Biometrika, 63, 581 592
14. Rubin D B (1987). Multiple Imputation for Nonresponse in Surveys. New York. John Wiley and Sons.
15. SAS Institute, 2005. Multiple Imputation for missing data. Concepts and New Approaches. A Useful Overview of the different methods to deal with Missing Data Using SAS
16. Schmitt p, Mandel J, Guedj M (2015) A Comparison of Six Methods of missng data Imputation. J. Biomet Biostat 6: 24 doi:10.4172/2155-6180.1000224
17. Schafer, J. L (197). Analaysis of incomplete Multivariate data. New York: Chapman & Hall
18. Sunbul Secil Omur(2018) The Impact of Diffrnt Missing Data handling Methods on DINA Model. Internatiopnal Journal of Evaluation and Research in education (IJER) Vol 7. No.1. pp 77 – 86 ISSN: 2252 – 8822
19. Therese D. pigott (2001) A Review of methods of Missing Data. Educational research and evaluation, Vol 7. No 4 pp. 353 – 388.